Importance Weighted Passive Learning

Shuaiqiang Wang
School of Computer Science and Technology
Shandong University of Finance and Economics
7366 2nd Ring Road East, Jinan 250014 China
swang@sdufe.edu.cn

Xiaoming Xi, Yilong Yin*
School of Computer Science and Technology
1500 Shunhua Road, Jinan 250101 China
Shandong University
fyzq10@126.com, ylyin@sdu.edu.cn

ABSTRACT

Importance weighted active learning (IWAL) introduces a weighting scheme to measure the importance of each instance for correcting the sampling bias of the probability distributions between training and test datasets. However, the weighting scheme of IWAL involves the distribution of the test data, which can be straightforwardly estimated in active learning by interactively querying users for labels of selected test instances, but difficult for conventional learning where there are no interactions with users, referred as *passive learning*. In this paper, we investigate the *insufficient sampling bias* problem, i.e., bias occurs only because of insufficient samples, but the sampling process is unbiased. In doing this, we present two assumptions on the sampling bias, based on which we propose a practical weighting scheme for the empirical loss function in conventional passive learning, and present IWPL, an importance weighted passive learning framework. Furthermore, we provide IWSVM, an importance weighted SVM for validation. Extensive experiments demonstrate significant advantages of IWSVM on benchmarks and synthetic datasets.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications–*Data mining*; I.2.6 [Artificial Intelligence]: Learning

General Terms: Algorithms, Performance, Experimentation

Keywords: Classification, Learning with confidence, Discounted confidence

1. INTRODUCTION

For correcting the bias of the probability distributions between training and test datasets, importance weighted learning [2, 7] was proposed, where the importance weight of each instance can be adjusted for the empirical loss function according to the probability distributions of the training and test datasets. However, the weighting scheme involves probability distribution of test dataset, which can be easily estimated in active learning by interactively querying users for labels of selected test instances, but difficult for conventional passive learning where there are no interactions with users. Besides, learning with confidence [4] can utilize another category of confidence labels for instances to generate the important weights. However, it requires much more efforts for labeling and the confidence scores might be inconsistent, for example, some users are self-confident and universally assign high confidences to instances while others are not and assign low confidences.

In this paper, we investigate the *insufficient sampling bias* problem, i.e., the bias occurs because of insufficient samples, but the sampling process is unbiased. For many sophisticated learning tasks such as speech recognition and information extraction, it is very difficult, time-consuming, or expensive to obtain sufficient labeled instances for training [6], and thus the training dataset may distribute differently from the test, and these datasets may be very noisy.

In doing this, we present two assumptions on the distributions of the training and test datasets for satisfying insufficient sampling bias, where the variances of the training and test instances can be different, however, their distribution categories and expectations should be the same. For example, the probability distributions of the training and test instances are two Gaussian distributions $\mathcal{N}(\mu, \sigma^2_{train})$ and $\mathcal{N}(\mu, \sigma^2_{test})$, which share a same expectation μ, but have different variances σ^2_{train} and σ^2_{test}. Obviously, our assumptions are weaker than i.i.d.

Based on the assumptions, we propose a practical weighting scheme for the empirical loss function in conventional passive learning, and present IWPL, an importance weighted passive learning framework. Furthermore, we provide IWSVM, an importance weighted SVM for validation.

Specifically, we evaluate the probability distribution of the training dataset with the dissimilarity between each instance to the distribution expectation of the instances. Besides, we estimate the distribution of the test dataset based on the ranks of the training instances if we order them according to their dissimilarity to the distribution expectation of the instances.

It is less precise but more robust to adopt the ranks of the objectives to approximately represent their probabilities, which have been explored for many problems. For example, in evaluation systems, Borda count[1] transforms the review scores of the items from different reviewers into a set of partial rankings, and assigns a score of $1/r_i$ to a given item t_i that has a rank of r_i, indicating that t_i has a probability of $1/r_i$ to be ordered at the first rank. Then these items can be aggregated and ordered according to their total scores for evaluation. This method can achieve acceptable performance even when the review data is very sparse.

The effectiveness of our approach can be understood from another perspective. In order to reduce the sample bias, the distribution of the instances should be discounted with a rank-based weight $\frac{1}{\log(1+r)}$. This rank-based weighting scheme is also adopted to dis-

[*] Corresponding author.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

count the relevance gain of the document at the rank i in the accuracy measure of NDCG [3] for document ranking in information retrieval, and has been widely used in XML retrieval, database and collaborative filtering [5].

2. PROBLEM STATEMENT

Conventional passive learning. Let $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$ be the training instances, where $x_i \in \mathcal{X}$ is a training instance following a probability distribution $P(x)$, and $y_i \in \mathcal{Y}$ is the label of x_i following a conditional probability distribution $P(y|x)$.

Let $f(x; \theta) : \mathcal{X} \rightarrow \mathcal{Y}$ be a classifier for estimating the class label y for the input instance x, where θ represents the parameters that need to learn. Let $L(y, f(x; \theta)) : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ be the loss function at an input instance x, which measures the discrepancy between the ground truth label value y and its estimation $f(x; \theta)$.

Since the instance X and the label Y are two random variables following the joint probability distribution $P(X, Y)$, a learning algorithm learns the parameter θ by minimizing the expectation of the loss function, formally:

$$L_{EXP} = \arg\min_{\theta} \left[\int_{\mathcal{X} \times \mathcal{Y}} P(x, y) L(y, f(x; \theta)) \, dxdy \right].$$

In conventional passive learning, a standard empirical loss function to learn the parameter θ is shown as follows:

$$L_{EMP} = \arg\min_{\theta} \left[\frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i; \theta)) \right].$$

However, conventional passive learning fails to consider importance weights for instance, resulting to deteriorated performances.

Importance weighted active learning. For many problems, the distribution of the training data can be quite different from the test data. Let $p_{train}(x)$ and $p_{test}(x)$ be the probability density functions corresponding to the distributions of the training and test data respectively. The importance of the instance x_i can be measured by the ratio of test and training densities, formally:

$$w_i = \frac{p_{test}(x_i)}{p_{train}(x_i)}.$$

In active learning, the probability distribution $p_{test}(x_i)$ at the instance x_i in the test dataset can be obtained via interactions with users. First of all, a small subset of unlabeled instances are selected from the test dataset to interactively query users for labels of selected test instances. Then the probability distribution of the test dataset can be estimated by responses of users.

In doing this, importance weighted active learning [2, 7] presents an unbias empirical loss function L_{IWAL}, where the function value $L_{IWAL}(x_i)$ at a given instance x_i was weighted by w_i, formally:

$$L_{IWAL} = \arg\min_{\theta} \left[\frac{1}{n} \sum_{i=1}^{n} w_i \, L(y_i, f(x_i; \theta)) \right].$$

Although importance weighted learning introduces a weighting scheme w_i to the empirical loss function, it involves probability distributions of test dataset, which is difficult for estimation in conventional passive learning. Generally this weighting scheme can be only successfully adopted in active learning, where the distribution of the test instances can be estimated by requesting to the users.

Learning with confidence. Learning with confidence [4] can utilize confidence scores, another category of labels besides the class labels, to evaluate the important weights for instances in the empirical loss function. In this situation, binary classification can be represented naturally as a regression problem.

However, first of all, it requires much more efforts for labeling. Even these confidence scores are available, they might be inconsistent, for example, some users are self-confident and universally assign high confidences to the instances while others are not and assign low confidences.

In summary, *to our best knowledge, there are no practical method to estimate importance weights of instances for conventional passive learning.*

Insufficient sampling bias. In this paper, we focus on the insufficient sampling bias, i.e., bias occurs only if we cannot obtain sufficient instances for training, but the sampling process is unbiased. In doing this, we present two assumptions on the training and test datasets for satisfying insufficient sample bias.

ASSUMEPTION 1. ***Category assumption.*** *The probability distributions of the training and test datasets (p_{train} and p_{test}) belong to a same distribution category.*

ASSUMEPTION 2. ***Expectation assumption.*** *For each class, the distribution expectations of training and test datasets are the same, formally:*

$$\mathbb{E}[X_{train}, Y_{train}] = \mathbb{E}[X_{test}, Y_{test}] = \mathbb{E}[X, Y],$$

where X_{train} and X_{train} be two random variables of the training and test instances respectively.

We can use the average value of the instances to estimate their expectation. For each class, the difference is very small between the average values of the training and test datasets, formally:

$$\forall y \in \mathcal{Y} : \left| \frac{1}{|\{(x_i^T, y)\}|} \sum_{\forall (x_i^T, y)} x_i^T - \frac{1}{|\{(x_j^E, y)\}|} \sum_{\forall (x_j^E, y)} x_j^E \right| \leq \epsilon,$$

where ϵ is a small positive real number, (x_i^T, y) and (x_j^E, y) are a training and a test instance respectively, and both of them are labeled as y.

In this case, the distribution variances of the training and test datasets can be different due to different sampling rounds, however, their distribution categories are the same, and their expectations are so close that they can be recognized as the same as well. For example, $p_{train} \sim \mathcal{N}(\mu, \sigma_{train}^2)$ and $p_{test} \sim \mathcal{N}(\mu, \sigma_{test}^2)$, two Gaussian distributions, are distribution density functions of the training and test datasets. They share a same expectation μ, but have different variances σ_{train}^2 and σ_{test}^2. Such relationship guarantees that the distribution of the test dataset can be estimated with the distribution of the training dataset.

Obviously, the category and expectation assumptions are weaker than the independent and identically distribution (i.i.d.) assumption. Thus the problem is: *based on these two assumptions, how to practically estimate importance weights of instances for conventional passive learning.*

3. IMPORTANCE WEIGHTED PASSIVE LEARNING

In this section, we introduce IWPL, an important weighted passive learning framework based on the category and expectation assumptions for the insufficient sampling bias problem.

3.1 Weighted Empirical Loss Function L_{IWPL}

Probability distribution of the training dataset. Let $\mathbb{E}(X)$ be the distribution expectation of the training instances, the distribution at a given training instance x_i involves two issues: the *distance* from x_i to $\mathbb{E}(X)$ and the *angle* of the corresponding vectors of the instances x_i and $\mathbb{E}(X)$.

Actually these two issues can also be estimated by the similarity/dissimilarity measures between instances. For example, the Euclidean distance measures the distance between instances, the vector cosine measures the angle of the corresponding vectors of the instances, or we can conceive a combined one that involves both.

Thus, it is very natural and straightforward to utilize the dissimilarity $dis\,(x_i, \mathbb{E}(X))$ between a given training instance x_i and the distribution expectation $\mathbb{E}(X)$ for estimating the probability distribution at x_i in the training dataset, formally:

$$\hat{p}_{train}(x_i) \propto \frac{1}{dis\,(x_i, \mathbb{E}(X))}.$$

Probability distribution of the test dataset. Assumptions (1) and (2) guarantee that the probability distribution of the training dataset in the training data can be used to estimate that of the test dataset.

The dissimilarities of two instances x_i and x_j to the distribution expectation, $dis(x_i, \mathbb{E}(X)) = 3$ and $dis(x_j, \mathbb{E}(X)) = 5$, involves two aspects of information: (i) the scores of 3 and 5, and (ii) their preference $dis(x_i, \mathbb{E}(X)) < dis(x_j, \mathbb{E}(X))$. For estimating the distribution of the test dataset with the training dataset, the scores are meaningless due to different variances in these two datasets, however, the preference still works.

EXAMPLE 1. *Let $\mathbb{E}(X)$ be the distribution expectation of the instances. Let x_i and x_j be two training instances following the probability distribution density function p_{train}, where $dis(x_i, \mathbb{E}(X)) < dis(x_j, \mathbb{E}(X))$ and $p_{train}(x_i) < p_{train}(x_j)$. Suppose that x_m and x_n are two test instances following the distribution density function p_{test} and $dis(x_m, \mathbb{E}(X)) < dis(x_n, \mathbb{E}(X))$. In this case, $p_{test}(x_m) < p_{test}(x_n)$ holds according to the category and expectation assumptions.*

If we order these instances based on their average dissimilarity to the distribution expectation, each instance receives a rank value. The importance weight of the training instances x_i should be proportionate to that of the test instance x_m if they have a same rank. The distribution of the test dataset is estimated formally as follows:

$$\hat{p}_{test}(x_i) \propto \frac{1}{\log(1 + r_i)}.$$

It is less precise but more robust to adopt the ranking of the objectives to approximately represent their probabilities, which have been explored for many problems. For example, in evaluation systems, Borda count[1] transformed the review scores of the items from different reviewers into a set of partial rankings, and aggregated them for group recommendation. This method can achieve acceptable performance even when the review data is very sparse.

Adopting the ranking of the objectives to approximately represent their probabilities has been explored for many problems. In the Borda count method [1] for ranking aggregation, the form of the $\frac{1}{r}$ has already been used to represent the rank-based gains of the votes (probabilities) for each reviewer. In $\hat{p}_{test}$ we use a logarithmic variant to yield a meaningful score for $\hat{p}_{test}(x_i)$ even if x_i has a very high rank.

The proposed loss function L_{IWPL}. With the estimations of the probability distributions of the training and test datasets, the importance weighting scheme can be estimated as follows:

$$\hat{w}_i \propto \frac{\hat{p}_{test}(x_i)}{\hat{p}_{train}(x_i)} = \frac{dis\,(x_i, \mathbb{E}(X))}{\log(1 + r_i)}.$$

Thus the loss function can be represented as follows:

$$L_{IWCL} = \arg\min_{\theta} \left[\frac{dis\,(x_i, \mathbb{E}(X))}{\log(1 + r_i)} \, L(y_i, f(x_i; \theta)) \right].$$

The effectiveness of our approach can be understood from another perspective. In order to reduce the sample bias, the distribution of the instances should be discounted with a rank-based weight $\frac{1}{\log(1+r)}$. This rank-based weighting scheme is also adopted to discount the relevance gain of the document at the rank i in the accuracy measure of NDCG [3] for document ranking in information retrieval, and has been widely used in XML retrieval, database and collaborative filtering [5].

3.2 Implementation of L_{IWPL} for SVM

In conventional SVM, the empirical loss function L_{EMP} is implemented by maximizing the margin among support vectors. Let ϕ be the function used to map the linear vector x_i into a higher (maybe infinite) dimensional space. Let $C > 0$ be the penalty parameter for classification errors. Let ξ_i be the slack variable. Let $K(x_i, x_j) = \phi(x_i^T x_j)$ be a kernel function. For binary classification problems, the maximum margin-based loss function for SVM is represented as follows:

$$\begin{aligned} \arg\min_{\theta,\xi} \quad & \frac{1}{2}||\theta||^2 + C \sum_{i=1}^{n} \xi_i \\ \text{s. t.} \quad & y_i(\theta^T \phi(x_i)) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

where $C \sum_{i=1}^{n} \xi_i$ refers to the hinge loss function, an implementation of the empirical loss function L_{EMP}, and $||\theta||^2$ refers to the regularization of the parameters for reduction of the classifier's complexity.

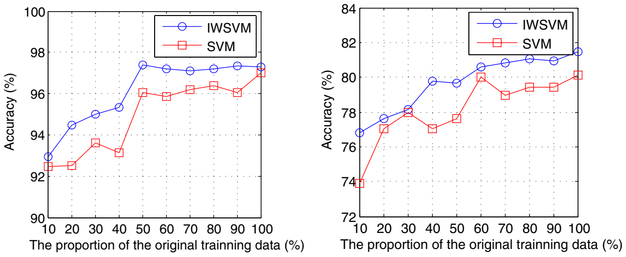
In our implementation of L_{IWPL}, the penalty of the classification error ξ_i is weighted by the estimation of the importance weight $\hat{w}_i$, formally:

$$\begin{aligned} \arg\min_{\theta,\xi} \quad & \frac{1}{2}||w||^2 + C \left(\sum_{y_i=+1} \hat{w}_i \, \xi_i + \sum_{y_i=-1} \hat{w}_i \, \xi_i \right) \\ \text{s. t.} \quad & \theta^T \phi(x_i)) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

4. EXPERIMENTS

We chose SVM as our comparison partner. Our implementation of IWSVM was based on SVM. A direct comparison of the two will provide valuable and irreplaceable insights.

We chose the liner kernel function in both IWSVM and SVM. In order to demonstrate the promises of IWSVM, we conducted two series of experiments on two benchmark and a synthetic datasets respectively. For each experiment, we ran each algorithm 5 times randomly and returned the average accuracy as the result for comparison.



(a) Accuracy for WDBC (b) Accuracy for Statlog (Heart)

Figure 1: Accuracy of the SVM and IWSVM with different numbers of the instances for training.

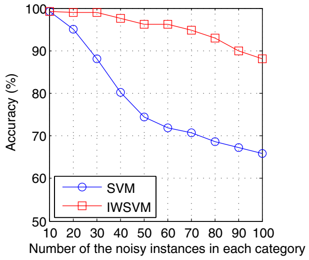


Figure 2: Accuracy of the SVM and IWSVM with different numbers of the noisy instances in each category.

4.1 Experiments with Small Data

This experiment demonstrates that IWSVM can achieve the same or even higher learning accuracy with a smaller dataset in comparison with conventional passive learning.

In this experiment, we used two classic UCI benchmarks, WDBC and Statlog (Heart). We randomly selecting different numbers of the instances from the training pool for IWSVM and SVM to evaluate their performance with small data. In particular, the proportions of the selected training instances varied from 10% to 100% in increments of 10%.

The experimental results are shown in Figure 1. From the figure we can see that IWSVM significantly gains in accuracy. Although the accuracy of the two methods improves with the increase of the the number of the training instances, IWSVM achieved the similar or even higher accuracy with a small volume data in comparison with SVM. In particular, for the WDBC dataset, as shown in Figure 1(a), IWSVM achieved the accuracy of 97.4% with only 50% training instances, while SVM could not outperform it even with 100% training instances. For the Statlog dataset, as shown in Figure 1(b), IWSVM achieved the accuracy of 80.6% with only 60% training instances, while SVM could not outperform it even with 100% training instances.

4.2 Experiments with Noisy Data

This experiment demonstrates the promises of the IWSVM with the noisy data in comparison with conventional passive learning.

We used a synthetic dataset for this experiment. For simplicity and easy demonstration, we employed two-dimensional features for representation of the instances. The generation process of the instances are as follows. First of all, we built two Gaussian models to generate the two categories of instances respectively. In particular, the first model generated positive instances, where the expectation and standard deviation were (0,0) and 0.5 respectively. The second model generated negative instances, where the expectation and standard deviation were (3,3) and 2 respectively. Every Gaussian model generated 500 instances for training and 500 instances for test. Then we gradually added the noisy instances into the original training data to evaluate the performances of IWSVM and SVM with noisy data. The positive noisy instances were generated by the Gaussian model for generating negative ones, and vice versa.

In this experiment, the numbers of the positive and negative noisy instances varied from 10 to 100 in increments of 10. The experimental results are shown in Figure 2. From the figure we can see that IWSVM significantly gains significantly in accuracy. Although the accuracy of the two methods deteriorate with the increase of the number of the noisy instances in each class, the accuracy of the IWSVM deteriorates much slower than that of the SVM with the increase of the number of the noisy instances. For example, both of the two methods achieved similar accuracies (a little more than 99%) with the original data; The learning accuracy of SVM were 74.4% and 65.9% when 50 and 100 noisy instances were introduced in each class, while the accuracy of IWSVM were 96.1% and 88.0% respectively.

5. CONCLUSION

For correcting the insufficient sampling bias between the probability distributions of the training and test datasets with conventional passive learning, we proposed a practical weighting scheme for estimating the importance of each instance, and presented IWPL, an importance weighted passive learning framework. Furthermore, we provide IWSVM, an importance weighted SVM for validation. We experimentally compared IWSVM with SVM on benchmarks and synthetic datasets, demonstrating the accuracy gain of IWSVM.

Given the novelty of the approach, for future work we plan to perform a systematic study on IWSVM, both theoretically and experimentally. Besides, it is interesting to explore the effectiveness of L_{IWPL} for other conventional passive learning algorithms such as logistic regression and neural network. Last but not least, it is essential to apply IWSVM to other application domains for validation purposes.

6. ACKNOWLEDGEMENT

This research was supported in part by the Natural Science Foundation of China (60970047, 61070097, 71171122), the Humanity and Social Science Foundation of Ministry of Education of China (12YJC630211), and the Natural Science Foundation of Shandong Province of China (BS2012DX012).

7. REFERENCES

[1] J. Aslam and M. Montague. Models for metasearch. In *SIGIR*, 2001.

[2] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst*, 20(4):422–446, 2002.

[4] M. Li and I. Sethi. Confidence-based classifier design. *Patt. Rec.*, 39(7):1230–1240, 2006.

[5] N. N. Liu and Q. Yang. EigenRank: A ranking-oriented approach to collaborative filtering. In *SIGIR*, 2008.

[6] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[7] M. Sugiyama, M. Krauledat, and K. Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8:985–1005, 2007.